



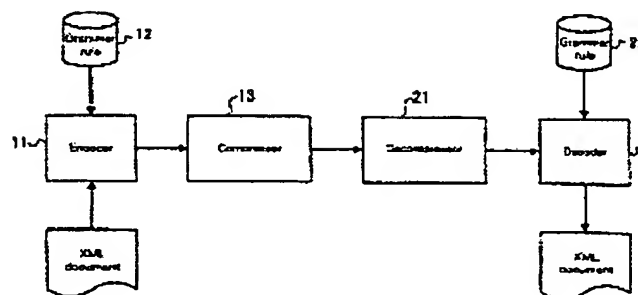
Data compaction, transmission, storage and program transmission**Patent number:** CN1316828**Publication date:** 2001-10-10**Inventor:** HIROSHI MARUYAMA (US); KENTO TAMURA (US);
NAOHIKO URAMOTO (US)**Applicant:** IBM (US)**Classification:****- international:** H03M7/30; G06F17/21**- european:****Application number:** CN20010103241 20010202**Priority number(s):** JP20000028359 20000204**Also published as:** EP1122655 (A2)
 JP2001217720 (A)

Report a data error he

Abstract not available for CN1316828

Abstract of corresponding document: EP1122655

A data compression apparatus for encoding data and for compressing the encoded data comprises: a grammar rule 12 for a tree local language in which data are represented by a labelled tree structure; an encoder 11 for reading a document written in the tree local language, for dividing the document into a structure part and contents, and for encoding the structure part using the grammar rule 12; and a compressor 13 for compressing the contents of the document extracted by the encoder 11, and for encoding the compressed contents.

**FIG. 1**

Data supplied from the esp@cenet database - Worldwide

[19] 中华人民共和国国家知识产权局

[51] Int. Cl⁷

H03M 7/30

G06F 17/21

[12] 发明专利申请公开说明书

[21] 申请号 01103241.3

[43] 公开日 2001 年 10 月 10 日

[11] 公开号 CN 1316828A

[22] 申请日 2001.2.2 [21] 申请号 01103241.3

[30] 优先权

[32] 2000.2.4 [33] JP [31] 28359/2000

[71] 申请人 国际商业机器公司

地址 美国纽约州

[72] 发明人 丸山宏 田村健人 浦本直彦

[74] 专利代理机构 中国专利代理(香港)有限公司

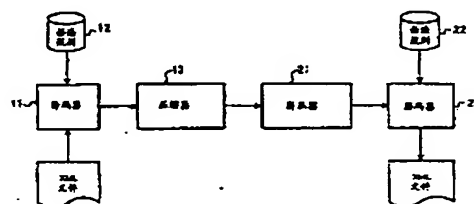
代理人 栾本生 傅康

权利要求书 4 页 说明书 12 页 附图页数 12 页

[54] 发明名称 数据压缩、传输、存储及程序传输

[57] 摘要

一种用于对数据编码和用于压缩编码数据的压缩设备包含:一个其中的数据是由标号树结构表示的树本机语言的语法规则 12;一个用于读取以该树本机语言编写的文件,把文件划分成结构部分和内容,和用语法规则 12 对该结构部分编码的编码器 11;一个用于压缩由编码器 11 提取的文件的内容,和对该压缩内容编码的压缩器 13。





权 利 要 求 书

1. 一种用于对数据编码和用于压缩该编码数据的压缩设备, 包含:
一个为其中的数据是由标号树结构表示的树本机语言存储语法规则的语法存储单元;

5 一个用于读取以该树本机语言编写的文件, 把文件划分成结构部分和内容, 并用语法存储单元中存储的语法规则对结构部分编码的编码器; 和

一个用于压缩由所述编码器抽出的所述文件的所述内容并对该压缩的内容进行编码的压缩器。

10 2. 按照权利要求 1 的数据压缩设备, 其中, 所述编码器包括:

一个用于将目标文件划分成结构部分和内容的划分器;

一个自动机构造器, 用于构造对应于所述语法规则的下推自动机;

15 一个编码数据生成器, 用于用由所述自动机构造器所构造的下推自动机来对由所述划分器获得的所述文件的所述结构部分进行语义分析, 并用于为该结构部分生成编码数据串。

3. 按照权利要求 2 的数据压缩设备, 其中, 所述编码器的所述编码数据生成器向在由所述自动机构造器所构造的所述下推自动机中驻留的选择分配符号, 并且, 所述编码数据生成器用所述下推自动机来分析以所述树本机语言编写的所述文件的所述结构部分, 并在选定的
20 各选择的位置, 输出为所述选择分配的符号, 以便为所述结构部分生成编码数据串。

4. 按照权利要求 1 的数据压缩设备, 其中, 所述压缩器不仅为以所述树本机语言编写的所述文件的所述内容, 也为由所述编码器获得的所述文件的所述结构部分, 进行压缩和编码。

25 5. 一种数据通讯系统, 包括:

一个用于在网络上发送数据的传输源数据处理设备; 和

一个用于接收由所述传输源数据处理设备在所述网络上发送的所述数据的传输目的地数据处理设备,

所述传输源数据处理设备包括:

30 一个用于为其中的数据是由标号树结构表示的树本机语言存储语法规则的第一语法存储单元,

一个用于读取以所述树本机语言编写的文件, 用于把所述文件划

分成结构部分和内容，和用于用所述第一语法存储单元中存储的所述语法规则对所述结构部分编码的编码器，

一个用于压缩由所述编码器提取的所述文件的所述内容并用于对该压缩内容编码的压缩器，和

- 5 一个用于发送由所述编码器编码的所述结构部分以及由所述压缩器压缩和编码的所述内容的发送器，并且

所述传输目的地数据处理设备包括：

一个用于从所述数据源数据处理设备接收数据的接收器，

- 10 一个用于存储与所述数据源数据处理设备的所述第一语法存储单元存储的所述语法规则相同的语法规则的第二语法存储单元，

一个用于采用与由所述数据源数据处理设备使用的压缩和编码方法对应的解压方法来解压由所述接收器接收的对应于所述文件的所述内容的数据的解压器，和

- 15 一个用于采用所述第二语法存储单元中存储的所述语法规则来解译由所述接收器接收的对应于所述文件的结构部分的数据的解码器。

6. 一种用于存储和管理存储单元中数据的数据库系统，包括：

一个为其中的数据是由标号树结构表示的树本机语言存储语法规则的语法存储单元；

- 20 一个用于读取以所述树本机语言编写的文件，把所述文件划分成结构部分和内容，并用所述语法存储单元中存储的所述语法规则对所述结构部分编码的编码器；

一个用于压缩由所述编码器提取的所述文件的所述内容并用于对该压缩内容编码的压缩器；

- 25 一个用于存储由所述编码器编码的所述文件的所述结构部分和存储由所述压缩器压缩和编码的所述文件的所述内容的存储单元。

7. 按照权利要求 6 的数据库系统，其中，所述压缩器不仅为以所述树本机语言编写的所述文件的所述内容，也为由所述编码器获得的所述文件的所述结构部分，进行压缩和编码。

- 30 8. 一种用于对数据编码和用于压缩编码数据的数据压缩方法，包括以下步骤：

读取以其中的数据是由标号树结构表示的树本机语言编写的文件，把所述文件划分成结构部分和内容；

用所述树本机语言的语法规则对所述结构部分编码;

压缩由所述编码器提取的所述文件的所述内容并用于对该压缩内容编码。

5 9. 按照权利要求 8 的数据压缩方法, 其中, 所述对所述文件的所述结构部分编码的步骤包括以下步骤:

构造对应于所述语法规则的下推自动机;

向在所述下推自动机中驻留的选择分配符号;

按照深度优先检索策略用所述下推自动机分析所述文件的所述结构部分, 并在所述选择的位置, 输出向所述选择分配的所述符号;

10 输出通过采用所述下推自动机而获得的符号串, 作为以所述树本机语言编写的所述文件的所述结构部分的编码数据串。

15 10. 按照权利要求 9 的数据压缩方法, 还包括: 一个在某属性属于所述树本机语言的某个目标文件时要在所述对以所述树本机语言编写的所述文件的所述结构部分编码的步骤之前执行的步骤, 即将所述属性改变为拥有所述属性的元素的子节点, 目的是将所述树本机语言的所述语法规则和所述文件转换成一个要由所述下推自动机处理的树结构。

20 11. 按照权利要求 8 的数据压缩方法, 还包括: 一个要在所述对所述文件的所述结构部分编码的步骤之后执行的步骤, 即采用另一个通用压缩和编码方法进一步对所述文件的所述编码结构部分进行压缩和编码。

12. 一种存储介质, 其上面的计算机输入装置存储一个计算机可读程序, 该程序允许计算机执行:

25 一个用于读取以其中的数据是由标号树结构表示的树本机语言编写的文件并且用于把所述文件划分成结构部分和内容的过程;

一个用于用所述树本机语言的所述语法规则对所述结构部分编码的过程; 和

一个用于压缩由所述编码器提取的所述文件的所述内容并且用于对该压缩内容编码的过程。

30 13. 一种程序传输设备, 包括:

用于存储程序的存储装置, 该程序允许计算机执行

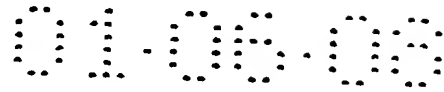
一个用于读取以其中的数据是由标号树结构表示的树本机语言编

写的文件并用于把所述文件划分成结构部分和内容的过程;

一个用于用所述树本机语言的语法规则对所述结构部分编码的过程;

5 一个用于压缩由所述编码器提取的所述文件的所述内容并且用于对该压缩内容编码的过程; 和

用于从所述存储装置读取所述程序并用于发送所述程序的传输装置。



说明书

数据压缩、传输、存储及程序传输

5 本发明涉及一种用于压缩以诸如 XML 或 ASN.1 的树本机语言 (tree local language) 编写的文件数据的数据压缩方法。

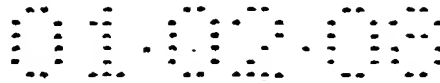
XML (eXtensible Markup Language-可扩充标记语言) 是用简单标签 (tags) 描述 (标记) 文件的逻辑结构的标记语言之一。在 XML 中, 为文件的成分规定语法规则, 提供逻辑定义, 使得用户能唯一地扩充文件数据。所以, 预计将来 XML 将作为一种数据格式, 用于因特网 10 网上的数据交换。

XML 有一个概念称为 DTD (文件类型定义), 可以确定某文件对特定 DTD 是否有效。举例来说, 规定一条语法规则, 使得节点<题目>、<作者>和<出版者>, 按照标着名称的顺序, 一次一个地出现在节点<书>的下面。可以确定预定的文件是否有效, 即预定的文件是否符合该语 15 法规则。

至于语言风格, XML 文件的结构属于一个称为树本机语言的类。按照树本机语言的定义, 数据是由标号 (labeled) 树结构表示的, 各个节点标号 (labels) 的正确数据是由子节点的标号的标准语言规定的。就是说, 在树本机语言中, 一个属于 (由 XML 中的 DTD 规定的) 20 某预定语法的树的集合, 是由指定各个节点的子节点的列表的标准语言确定的。这个类型的另一个树本机语言是 ASN.1 (Abstract Syntax Notation 1-抽象语义表示法 1)。

采用 XML 为商业应用和其它目的编写以前用 HTML 或网络上可用的其它资源不能编写的复杂数据结构, 有上升的趋势。预计, 有了这个 25 系统, 大型 XML 文件将有可能被应用程序交换。

一般来说, 为了交换数据或在数据库中存储的数据, 要对数据文件进行压缩, 以降低文件的大小, 提高传输效率。因此, 许多可用于各种数据格式类型的通用数据压缩技术和许多只适合特定数据格式类型的专用数据压缩技术, 已经被开发出来, 可用于 XML 文件的交换。 30 就压缩而言, 应当注意到, 尽管考虑了 XML 文件的数据结构, 也有对标签的明显多余的使用, 以便能指望有高的压缩比例。



如上所述，为方便数据的交换和在数据库中存储数据，一般要进行对数据文件的压缩。用树本机语言，如 XML，可以期望对数据部分 - 如代表文件结构的标签 - 有高的压缩比例。

5 假设对于数据通讯来说，双方采用共同的语法 G 并保证它们只交换对该语法有效的 XML 文件。还假设对语法规则的规定，使得在节点<题目>、<作者>和<出版者>，按照标着名称的顺序，一次一个地出现在节点<书>的下面。在这种情况下，当 XML 文件的接收者例如在 XML 文件中发现标签<书>时，该接收者就能预测到第一个子节点将是<题目>。因此，根据这个假设，标签<题目>变得多余，于是可以设计一种对 XML
10 文件编码的方法用以优化消息长度。这不仅适用于 XML，也适用于任意树本机语言（例如 ASN.1）。

然而按照惯例，用压缩来对树本机语言的文件的结构部分进行编码的压缩方法尚未提出。

所以，本发明的一个目的是采用数据压缩来对诸如 XML 或 ASN.1
15 的树本机语言编写的文件的结构部分进行编码。

本发明的另一个目的是为树本机语言提供一种与另一个通用数据压缩技术一起采用的专用数据压缩方法，以便能实现高的压缩比例。

为了实现以上目的，按照本发明，一种用于对数据编码和用于压缩该编码数据的数据压缩设备包含：一个为其中的数据是由标号树结构表示的树本机语言存储语法规则的语法存储单元；一个用于读取以
20 该树本机语言编写的文件，把该文件划分成结构部分和内容，并用语法存储单元中存储的语法规则对该结构部分进行编码的编码器；和一个用于压缩由编码器提取的文件的内容并用于对该压缩内容编码的压缩器。树本机语言是一种树语言，其中的数据是由标号树结构表示的，
25 并且其中，对于各个节点标号，采用用于子节点标号的标准语言来规定正确的数据。

该编码器包括：一个用于将目标文件划分成结构部分和内容的划分器；一个自动机构造器，用于构造对应于该语法规则的下推自动机；一个编码数据生成器，采用由自动机构造器所构造的下推自动机来对
30 由划分器获得的文件的结构部分进行语义分析，并用于为该结构部分生成编码数据串。

该编码器的编码数据生成器向在由自动机构造器所构造的下推自

动机中驻留的选择分配符号。该编码数据生成器采用下推自动机来分析以树本机语言编写的该文件的结构部分，并在选定的各选择的位置，输出为这些选择分配的符号，以便为该结构部分生成编码数据串。用这个方案，就能将用诸如标签的标号编写的文件的结构部分改变（编码）成一个简单的编码序列。为了用下推自动机分析文件结构部分，按深度优先检索策略跟踪文件的树结构。就是说，不是用与父节点等距离的次序，而是采用沿深度方向的节点之间的关系（父子关系）作为优先次序跟踪树，进行分析。

此外，压缩器不仅为以树本机语言编写的文件的内容，也为由编码器获得的文件的结构部分，进行压缩和编码。尽管由压缩器使用的压缩方法没有特别的限制，可以采用的普通的通用方法。如果编码器通过对文件的结构部分的编码获得标准数据串，该编码器可采用诸如PKZIP的通用方法，为编码数据串执行压缩和编码，预期能有高的压缩比例。因此，在文件的内容被压缩时，最好也压缩该编码数据串。

还有，如果将结构部分的编码数据串与文件的内容结合起来，然后将产生的数据压缩，该结构部分和该内容就构成一个单一的文件。这对文件管理来说更可取。

按照本发明，一种数据通讯系统包含：一个用于在网络上发送数据的传输源数据处理设备；一个用于接收由传输源数据处理设备在网络上发送的数据的传输目的地数据处理设备。该传输源数据处理设备包含一个用于为其中的数据是由标号树结构表示的树本机语言存储语法规则的第一语法存储单元，一个用于读取以该树本机语言编写的文件、用于把文件划分成结构部分和内容并用于采用该第一语法存储单元中存储的语法规则对该结构部分进行编码的编码器，一个用于压缩由编码器提取的文件的内容并用于对该压缩内容编码的压缩器，以及一个用于发送由编码器编码的结构部分以及由压缩器压缩和编码的内容的发送器。该传输目的地数据处理设备包含一个用于从数据源数据处理设备接收数据的接收器，一个用于存储与数据源数据处理设备的第一语法存储单元存储的语法规则相同的语法规则的第二语法存储单元，一个采用与由数据源数据处理设备使用的压缩和编码方法对应的解压方法来解压由接收器接收的对应于文件的内容的数据的解压器，以及一个用于采用第二语法存储单元中存储的语法规则来解译由接收

器接收的对应于文件的结构部分的数据的译码器。实现该过程的一个高级方法，是让所准备的语法规则由数据传输源和目的地共同使用，因为对于数据通讯来说，可以为以树本机语言编写的文件获得高的压缩比例，并且能提高通讯效率。因为对于商业通讯来说，一般法则是把树本机语言的语法规则共同使用，所以能容易引用本发明。

此外，按照本发明，一种用于存储和管理存储单元中的数据的数据数据库系统包含：一个为其中的数据是由标号树结构表示的树本机语言存储语法规则的语法存储单元；一个用于读取以该树本机语言编写的文件，把文件划分成结构部分和内容，并用语法存储单元中存储的语法规则对结构部分编码的编码器；一个用于压缩由编码器提取的文件的内容并用于对该压缩内容编码的压缩器；一个用于存储由编码器编码的文件的结构部分和存储由压缩器压缩和编码的文件的内容的存储单元。

该压缩器不仅为以树本机语言编写的文件的内容，也为由编码器获得的文件的结构部分，进行压缩和编码。如果将结构部分的编码数据串与文件的内容结合起来，然后将产生的数据压缩，结构部分和内容就构成一个单一的文件；对文件管理来说，这更可取。

按照本发明，一种用于对数据编码和用于压缩编码数据的数据压缩方法包含以下步骤：读取以其中的数据是由标号树结构表示的树本机语言编写的文件，并把该文件划分成结构部分和内容；用该树本机语言的语法规则对结构部分进行编码；压缩由编码器提取的文件的内容并用于对该压缩内容进行编码。

对文件的结构部分编码的步骤包括步骤：构造对应于该语法规则的下推自动机；向在下推自动机中驻留的选择分配符号；按照深度优先检索策略用下推自动机分析该文件的结构部分，并在各选择的位置，输出向这些选择分配的符号；输出通过采用下推自动机而获得的符号串，作为以树本机语言编写的文件的结构部分的编码数据串。用这个方案，就能将用诸如标签的标号编写的文件的结构部分改变（编码），获得一个简单的编码序列。

该数据压缩方法进一步包括：一个当某属性属于树本机语言中某个目标文件的节点时要在对以树本机语言编写的文件的结构部分编码的步骤之前执行的步骤，即将该属性改变为拥有该属性的元素的子节

点，以便将树本机语言的语法规则和文件转换成一个要由下推自动机处理的树结构。这个方案之所以更可取，是因为即使该属性包含在如 XML 文件的目标文件中，也能用下推自动机对该结构部分进行编码。

5 该数据压缩方法还包含：一个要在对文件的结构部分编码的步骤之后执行的步骤，即采用另一个通用压缩和编码方法进一步对编码的文件结构部分进行压缩和编码。这个方案之所以更可取，是因为预期能有更高的压缩比例。

10 按照本发明，提供一种存储介质，其上面的计算机输入装置存储一个计算机可读程序，该程序允许计算机执行：一个用于读取以其中的数据是由标号树结构表示的树本机语言编写的文件并且用于把该文件划分成结构部分和内容的过程；一个采用该树本机语言的语法规则对结构部分编码的过程；一个用于压缩由编码器提取的文件的内容并且用于对该压缩内容编码的过程。用这个方案，所有其中安装这个程序的信息处理设备都能在压缩以该树本机语言编写的文件时实现高的
15 压缩比例，并且能获得较高的通讯和存储效率。

此外，按照本发明，一种程序传输设备包含：用于存储程序的存储装置，该程序允许计算机执行一个用于读取以其中的数据是由标号树结构表示的树本机语言编写的文件并用于把该文件划分成结构部分和内容的过程，一个采用该树本机语言的语法规则对该结构部分编码
20 的过程，一个用于压缩由编码器提取的文件的内容并且用于对该压缩内容编码的过程；用于从该存储装置读取该程序并用于发送该程序的传输装置。用这个方案，所有已经从该程序传输设备下载这个程序并安装该程序的信息处理设备，都能在压缩以该树本机语言编写的文件时实现高的压缩比例，并且能获得较高的通讯和存储效率。

25 图 1 是解释按照本发明一个实施例的文件压缩系统的总体方案的示意图；

图 2 是解释按照实施例的数据压缩处理的示意图；

图 3 是解释按照实施例的编码器的示意图；

图 4 是表示按照实施例的一例目标 XML 文件的示意图；

30 图 5 是表示图 4 中的 XML 文件结构部分的示意图；

图 6 是表示用于实施例的一例语法规则的示意图；

图 7 是表示按照图 6 中的语法规则构造的下推自动机的示意图；

图 8 是表示用于解释使用下推自动机的语法检查方法的一例语义树的示意图;

图 9 是通过使用图 7 中下推自动机而生成的编码转换器 (transducer) 的示意图;

5 图 10 是表示通过为图 8 中的语义树进行有效性检查而获得的结果例子的示意图;

图 11 是通过使用图 7 中下推自动机而生成的解码转换器的示意图;

10 图 12 是解释其中将有属性的 DTD 转换成无属性的 DTD 的状态的示意图;

图 13 是解释其中将有属性的 XML 文件转换成无属性的 XML 文件的状态的示意图;

图 14 是解释对数据通讯系统应用实施例时的结构的示意图;

图 15 是解释对数据库系统应用实施例时的结构的示意图;

15 现在将参考附图, 详细地说明本发明的最佳实施例。

图 1 是解释按照本发明的一个文件压缩系统的总体方案的示意图。图 1 中, 编码器 11 将目标文件划分成结构部分和内容, 并用在预定存储器中存储的语法规则 12 对结构部分编码。压缩器 13 包含由编码器 11 编码的结构部分和文件的内容。解压器 21 解压由压缩器 13 压缩的文件。在文件被解压器 21 解压的时候, 文件被分离成内容和由编码器 11 编码的结构部分。解码器 23 通过使用在预定存储器中存储的语法规则 22, 重新构造编码的结构部分, 将结构部分与内容结合起来, 重新产生文件。

25 当将实施例的方法用于数据通讯的数据压缩时, 将编码器 11 和压缩器 13 配置在发送端, 将解压器 21 和解码器 23 配置在接收端。当将实施例的方法用于压缩要在数据库系统中存储的数据文件时, 按照数据发送, 编码器 11 起译码器 23 的作用, 压缩器 13 起解压器 21 的作用。

现在将就用 XML 作为目标树本机语言的例子给出解释。

30 图 2 是解释按照实施例的数据压缩处理的示意图。在图 2 中的数据压缩处理中, 首先, 目标 XML 文件 201 被从编码器 11 读出, 划分成结构部分 202 和内容 204。结构部分 202 包括 XML 文件的树结构、

标签名和属性名；内容 204 包含 #PCDATA 和 XML 文件的属性值。之所以将 XML 文件划分成结构部分 202 和内容 204，是因为一般来说结构部分 202 和内容 204 有相当不同的统计偏差，独立地压缩这二者是效率高的。

5 将通过划分 XML 文件 201 所获得的结构部分 202 用编码器 11 进行编码，并且将语法规则 12 用于这个编码。由于在本实施例中 XML 文件是目标，语法规则 12 由 DTD 规定。这个编码处理将在后文作详细说明。所获得的编码数据串 203 和内容 204 被传送到压缩器 13。

10 最后，压缩器 13 对编码数据串 203 和内容 204 进行压缩和编码，将获得的数据组合起来，生成压缩 XML 文件 205。为了进行该编码过程，压缩器 13 采用常规的有名方法，诸如 LZ77。此时，压缩器 13 主要被用来对内容 204 进行压缩和编码。然而，对编码数据串 203 可以有效地使用通用压缩与编码方法，诸如 PKZIP。正如后文中将要说明的那样，在本实施例中，编码数据串 203 是作为数字序列被输出的。
15 因此，如果数据是一序列的比较规则的数字，就可以期望有高的压缩比例。所以，压缩器 13 可以将编码数据串 203 与内容 204 一起进行压缩和编码。不过应当注意，压缩器 13 对编码数据串 203 的压缩是个任意过程。编码数据串 203 和内容 204 可以不由压缩器 13 压缩，而可以只是被彼此关联或结合起来，可以交换或存储在存储器中。

20 如上所述，在本实施例中，XML 文件的结构部分 202 是用本发明方法压缩的，此外，编码结构部分 202 和内容 204 是用常规方法压缩的。所以，本发明方法是与各种常规压缩方法结合起来使用的。

为了将这样压缩的 XML 文件解压，要反过来进行以上的压缩处理。具体来说，解压器 21 用与压缩器 13 所使用的压缩和编码方法相对应
25 的方法对编码数据串 203 解压。然后，如下文将要详细说明的那样，解码器 23 用语法规则 22 重新构造被解压的编码数据串 203。语法规则 22 与语法规则 12 相同，是由 DTD 规定的。然后，用在解码过程中获得的结构部分 202 和由解压器 21 解压的内容 204 重新生成 XML 文件 201。

30 现在将详细解释按照本实施例进行的用于对 XML 文件的结构部分编码的处理。

为简化解释，对于这个过程来说，目标 XML 文件不含任何属性，XML

文件的实际总体是设计好的。如何处理属性将在以后作讨论。

图 3 是解释对 XML 文件的结构部分编码的编码器 11 的方案的功能框图。在图 3 中, 编码器 11 包含: 划分器 111, 用于将目标 XML 文件 201 划分成结构部分 202 和内容 204; 自动机构造器 112, 用于用语法规则 12 构造将在以后作说明的下推自动机; 编码数据串生成器 113, 用于通过用由自动机构造器 112 构造的下推自动机作为编码转换器, 为结构部分 202 生成编码数据串 203。

图 4 是表示一例目标 XML 文件的示意图。XML 文件的内容由字符串的列表组成, 字符串位于对应于 #PCDATA 的内容模型 (contents model) 的部分。就是说, 图 4 中的 XML 文件的内容, 是一个由 4 个字符串 “String1”、“String2”、“String3”、和 “String4” 组成的列表。该列表例如可以用下列字节串来紧凑地表示, 该字节串中, 将各字符串以空字符作为结束 (“\0” 代表空字符)。

“String1\0String2\0String3\0String4\0”

如上所述地将这个字符串独立于结构部分进行压缩和编码。

图 4 中的 XML 文件结构部分在图 5 中表示。这个结构部分是通过将对应于图 4 中 XML 文件的内容的字符串替换为占位符 (□) 而获得的。

本实施例中, 译码器 11 的划分器 111 从图 4 中的 XML 文件提取图 5 中的结构部分, 自动机构造器 112 用语法规则 12 构造下推自动机, 编码数据串生成器 113 用下推自动机对结构部分编码。图 6 是表示用于规定语法规则 12 一例 DTD 的示意图。

在划分器 111 执行了划分过程之后, 为了用语法规则 12 进行编码, 自动机构造器 112 构造对应于 DTD 的下推自动机。按照图 6 中的 DTD, 当元素 A 出现时, 意味着元素 B 和元素 C 将按照标着名称的顺序逐一地出现, 该状态转换 (state shifting) 然后结束。类似地, 当元素 B 出现时, 意味着元素 D 将出现, 该状态转换然后结束。当元素 C 出现时, 意味着 0 个元素或元素 E 或元素 F 将出现, 该状态转换然后结束。当元素 E 出现时, 意味着一个元素 G 或一个元素 H 将出现, 该状态转换然后结束。

图 7 是表示对应于图 6 中的 DTD 的自然下推自动机的示意图。由于非结束符 (non-terminal symbol) D 和 G 是明显的只有结束符 #

PCDATA 的规则，它们没有在图中显示。

可以为语法的各个非结束符构造没有二义的自动机。因此，如果将本实施例应用于数据通讯，可以通过利用发送端和接收端公用的 DTD 来构造同样的下推自动机。

- 5 一般来说，下推自动机被用来分析输入串的语义。在这个意义上，下推自动机接收表面层上的所有符号串，即所用通过设置一个或多个 #PCDATA（或者占位符“□”）而获得的串。然而例如，作为获得的语义分析数，节点 B 和节点 C 必须作为节点 A 的子节点按照标着名称的顺序出现。此外，紧接元素 C 之后，空状态被转换到最后状态。如上
10 所述，下推自动机可以被用来确定被分析的 XML 文件的语义分析树是否满足语法。

- 现在将采用图 8 中的语义树作为例子，解释用下推自动机检查语法所进行的处理。在图 8 中，没有显示每个树叶上的 #PCDATA。为了
15 确定这个语义树是否能由图 6 中的 DTD 规定的语法生成，只需要语义树的每个节点能确定由其子节点组成的串是否能被对应于该节点的非结束符接收。例如，元素 A 的子节点是串 BC。这些子节点被对应于非结束符 A 的自动机（见图 7 中的 A）接收。因此，发现这个部分满足语法。如果通过使用对应自动机按预排序以同样的方式（按深度优先检索策略跟踪）遍历所有的节点，语法检查就结束。

- 20 下推自动机对语义分析树的这个用法，在以下的解释中被称为有效性检查。应当注意，除了 ϵ 到最终状态的转换之外，对应于以上过程中使用的每个非结束符的自动机都是最小决策自动机（minimum decisive automata）。

- 25 自动机构造器 112 将图 7 中的下推自动机转换成一个对 XML 文件的结构部分（见图 5）编码的转换器，即用于分析字符串的语义的自动机。

- 在图 7 中的下推自动机中，假设某项是一个由 4 个 #PCDATA（或占位符“□”），并且分析过程是在用 A 作为开始符时启动的。然后，顺序地生成节点 A、节点 B、节点 C 和节点 D，识别第一个 #PCDATA。
30 节点 C 被生成时，有三个选择：可以生成节点 E，可以生成节点 F，或者状态转换可以在节点 C 被结束并返回到上层节点。将数字 1、2 和 3 按标签的字母顺序分配给这三个选择（标签 ϵ 总是被确定是最后的）。

类似地，因为节点 E 的第一个状态提供两种选择，或者可以生成节点 G，或者可以生成节点 H，所以将数字 1 和 2 分配给这些选择。在本实施例中，分配给选择的是数字，但是可用来标识选择的符号并不限于数字。任意符号，如字母字符或符号，都可以用来表明选择。

5 图 9 是通过转换图 7 中的下推自动机而生成的编码转换器的示意图。

编码器 11 的编码数据串生成器 113 运行由自动机构造器 112 构造的编码转换器。

10 进行有效性检查（按先根次序应用规则）时，图 9 中的编码转换器输出对应的选择号。具体来说，在图 9 中，没有对应规则 A、B、F 和 H 的选择，编码转换器没有输出。然而，当规则 C 和规则 E 被使用时，编码转换器输出适当的号码。例如，当对图 8 中的语义树进行有效性检查时，编码转换器在跟踪该树时，输出图 10 中所示的号码。

15 通过以上处理，获得号码串“112123”，它严格地规定下推自动机的运动。因此，该号码串可以被用作图 4 中的 XML 文件的结构部分（图 5）的编码数据串。

现在将说明按照本实施例解译 XML 文件的结构部分的处理。

20 要解译通过以上处理编码的 XML 文件，只需要将编码转换器的输入/输出反过来。因此，译码器 23 用与图 7 中的同样的下推自动机来生成解码转换器，开始解码过程。如上所述，因为可以为语法的各个非结束符构造没有二义的自动机，如果由 DTD 规定的语法规则 12 与语法规则 22 相同，译码器 23 就能构造出与图 7 中的完全相同的下推自动机。

25 图 11 是通过转换与图 7 中的相同的下推自动机而生成的解码转换器的示意图。在图 11 中的解码转换器中，“i/B”代表“当输入字符串“I”出现时，调用规则 B，然后将状态转换到下一个”的转换。这样，从译码器 11 输入一个号码串，生成一个对应的语义分析树。

30 如果根据原始号码的分配，输入上述号码串“112123”，下推自动机（译码转换器）就能没有二义地接受 XML 文件的编码号码串。因此，所生成的语义分析树与图 8 中的原始语义分析树相同。结果，就能重新生成 XML 文件的结构部分。

现在将说明对属性的处理。

在本实施例中，将属性转换成树结构，以便能由下推自动机作处理。具体来说，改变所有有属性的元素（ELEMENT）改变，使得将属性看作子节点。此时，属性以它们名称的字母顺序出现。让属性 # REQUIRED（# 必需的）不变，并让属性 # IMPLIED（# 隐含的）带有选项“？”。因为起初没有为属性 # FIXED（# 固定的）提供信息，所以它不包含在通过转换而获得的 DTD 中。

图 12 是表示预定的 DTD 在换成前与转换后的状态的比较的示意图。图 12 中，左边的 DTD 被转换成右边所示的形式。图 13 是表示预定的 XML 文件在换成前与转换后的状态的比较的示意图。

按以上方式将 DTD 和 XML 文件改变到没有提供属性的状态，就执行了上述的编码和译码处理。应当注意的是，DTD 的转换可以在下推自动机的构造之前提前进行，或者可以在下推自动机已经被构造之后按需进行。在第一种情况下，由转换所得的新 DTD 被用来构造下推自动机。在第二种情况下，原始 DTD（有属性的）被用来构造下推自动机。

如上所述，按照本实施例，XML 文件压缩端和解压端不可避免地要使用共同的相同 DTD。因此，如果将本实施例的数据压缩方法用于数据通讯，就必须为发送端数据处理设备和接收端数据处理设备准备相同的 DTD。

图 14 是解释采用本实施例的数据通讯系统的配置的示意图。在发送端的数据处理设备 1410 中，译码器 11 接收来自数据处理器 XML 文件，并用（对应于图 1 中的语法规则 12 的）DTD 1411 来对结构部分编码。压缩器 13 压缩编码结构部分和内容，发送器 1412 通过通讯网络向接收端发送数据处理设备 1410 中的由编码器 11 进行的编码和压缩器 13 进行的压缩而生成的结果 XML 文件。在接收端的数据处理设备 1420 中，接收器 1422 通过通讯网络接收数据，并将它们发送到解压器 21。此时，解压器 21 解压所接收的数据，将 XML 文件的内容恢复。译码器 23 然后用（对应于图 1 中的语法规则 12 的）DTD 1421 来解译已经被解压的数据的结构部分的编码数据串。译码器 23 然后重新装配所获得的结构部分和内容，以重新生成 XML 文件，并将该 XML 文件发送给数据处理器。在这个处理期间，发送端的数据处理设备 1410 中的 DTD1411 与接收端的数据处理设备 1420 中的 DTD1421 有相同的

内容。

如果 XML 文件是由用于商业通讯的应用交换的,例如是在电子商务事务处理期间交换的,在大多数情况下,要经互相同意而预先建立一个 DTD。因此,假设 DTD 将被共同使用,则本实施例可以应用于商业通讯。

当用本实施例的方法来压缩要由数据库系统存储的数据文件时,要解译 XML 文件的结构部分,可以原封不动地用对结构部分编码所使用的 DTD 来解译该结构部分,这样就不必考虑是否要共同使用一个 DTD。

图 15 是解释采用本实施例的数据库系统的配置的示意图。在数据库系统 1500 中,译码器 11 接收来自数据处理器的 XML 文件,并用(对应于图 1 中的语法规则 12 的)DTD 1501 来对结构部分编码。压缩器 13 然后压缩编码结构部分和内容。XML 文件被译码器 11 编码并被压缩器 13 压缩后,被存储在存储器 1502 中。要从存储器 1502 读取 XML 文件,压缩器 13 起着解压器 21 的作用,编码器 11 起着译码器 23 的作用,用于对 XML 文件的结构部分编码的 DTD 1501 被用于解译该结构部分。

在上述解释中,用 XML 语言作为树本机语言。然而,本实施例可用于另一个树本机语言,如 ASN.1。不过在这种情况下,语法规则如 XML 的 DTD,也必须由数据文件压缩端和解压端共同使用。

如上所述,按照本发明,可以通过进行数据压缩来对树本机语言的文件的结构部分编码。

此外,因为数据压缩方法特别适合于树本机语言并且是与另一个通用数据压缩技术一起使用的,所以可以采用一个提供高压缩比例的数据压缩方法。

说明书附图

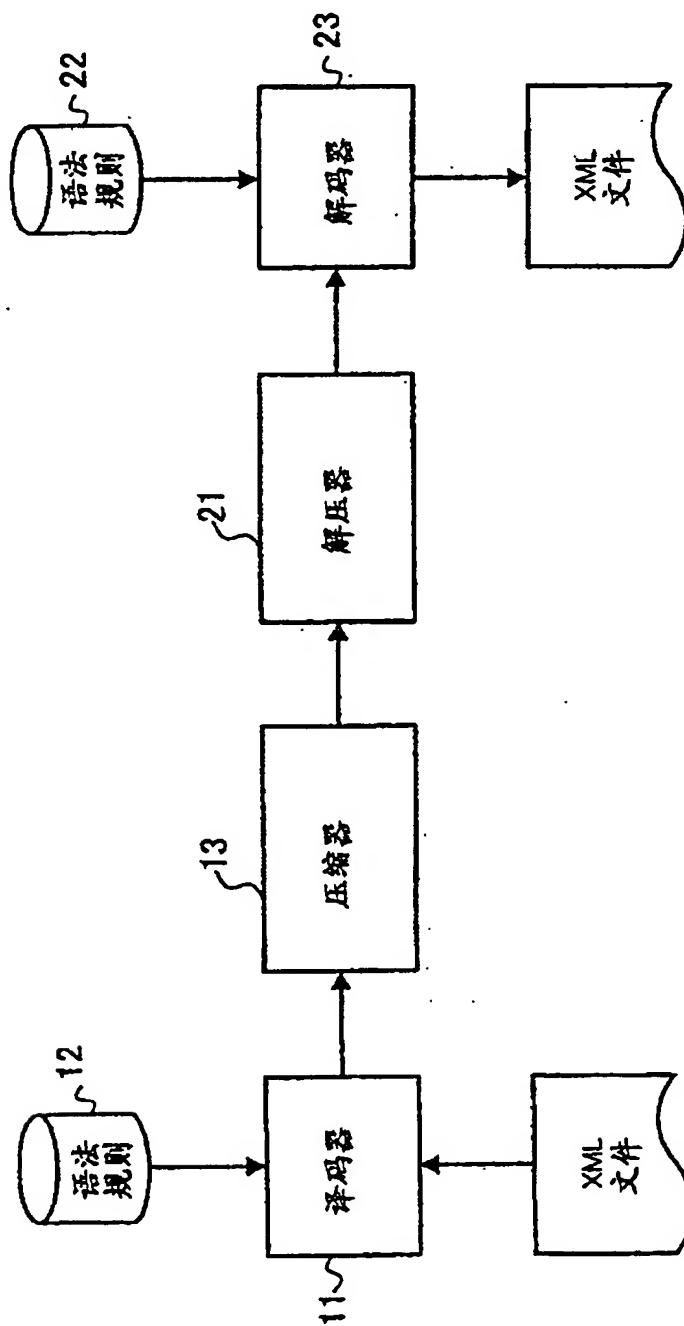


图 1

01:02:03

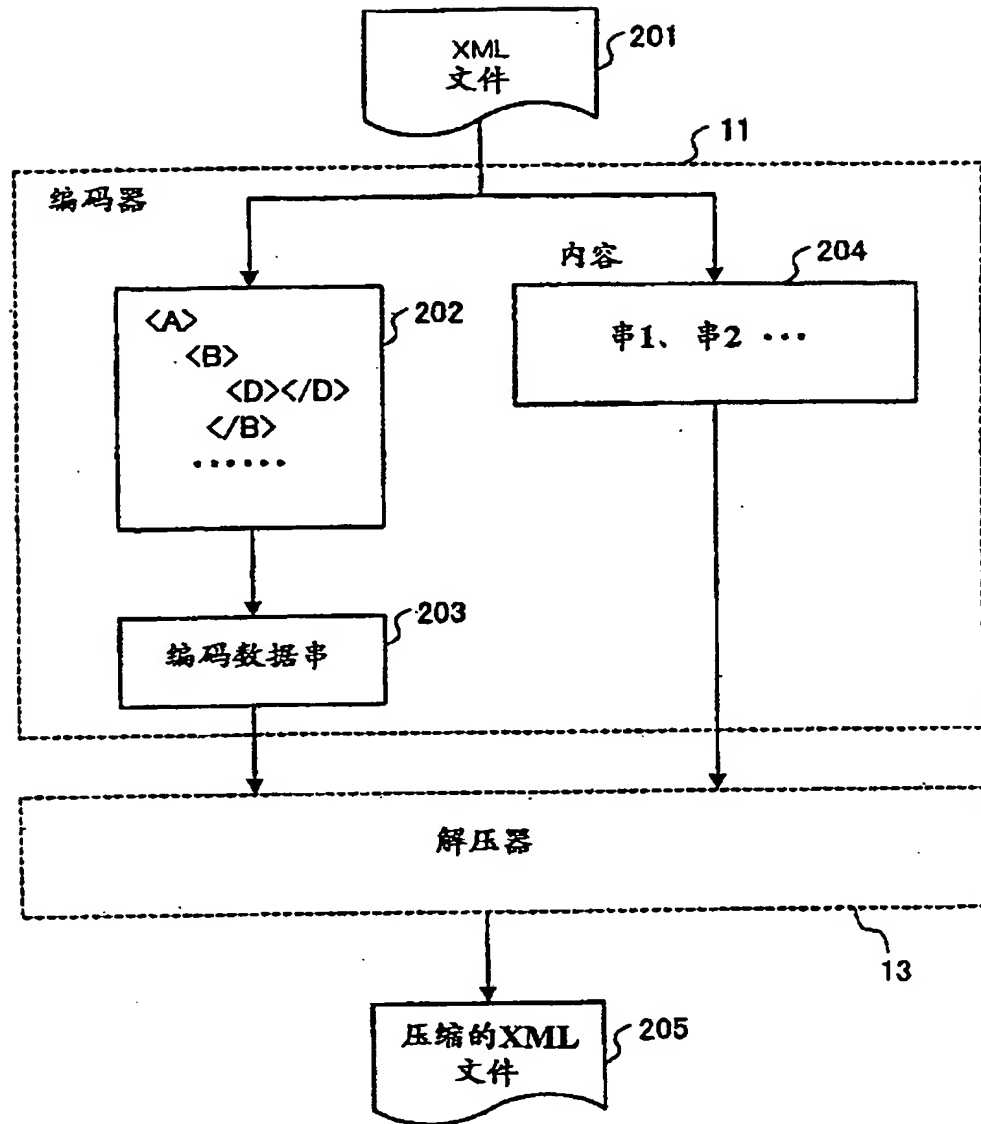


图 2

01.02.08

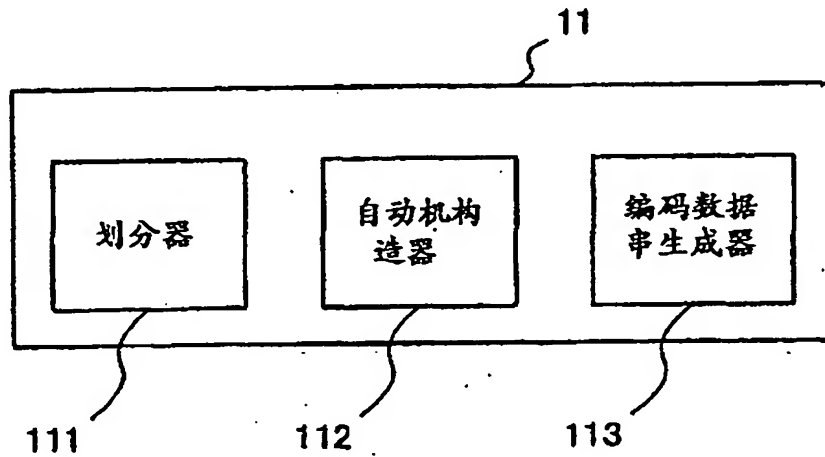


图 3

```
<?xml version="1.0"?>
<!DOCTYPE department SYSTEM "sample.dtd">
<A>
  <B>
    <D>String1</D>
  </B>
  <C>
    <E>
      <G>String2</G>
    </E>
    <F>
      <G>String3</G>
    </F>
    <E>
      <H>
        <G>String4</G>
      </H>
    </E>
  </C>
</A>
```

图 4

01.02.08

```
<?xml version="1.0"?>
<!DOCTYPE department SYSTEM "sample.dtd">
<A>
  <B>
    <D>□</D>
  </B>
  <C>
    <E>
      <G>□</G>
    </E>
    <F>
      <G>□</G>
    </F>
    <E>
      <H>
        <G>□</G>
      </H>
    </E>
  </C>
</A>
```

图 5

```
<ELEMENT A (B, C)>
<ELEMENT B (D)>
<ELEMENT C (E|F)*>
<ELEMENT E (G|H)>
<ELEMENT F (G)>
<ELEMENT H (G)>
<ELEMENT D (#PCDATA)>
<ELEMENT G (#PCDATA)>
```

图 6

01.02.08

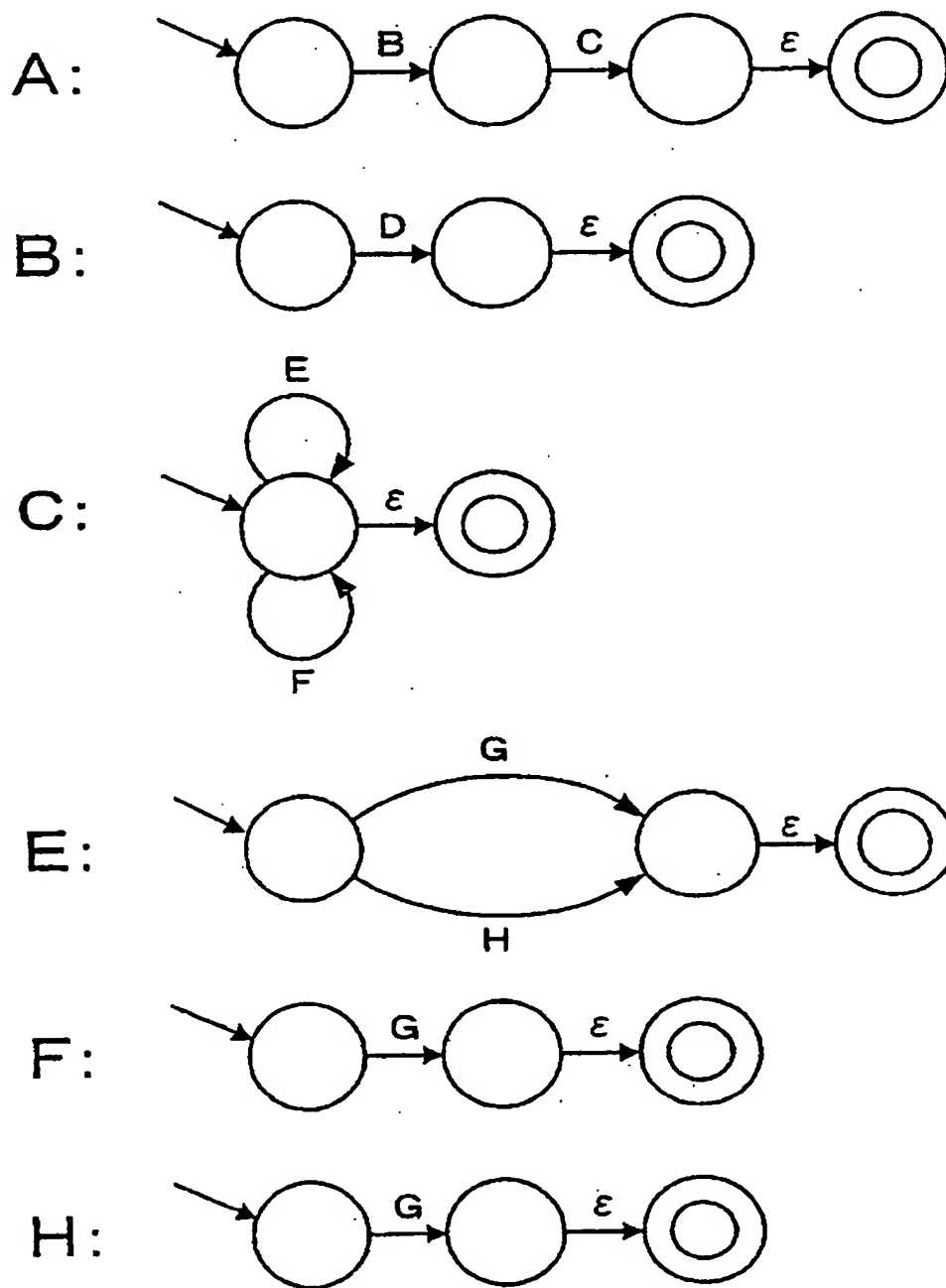


图 7

01.02.08

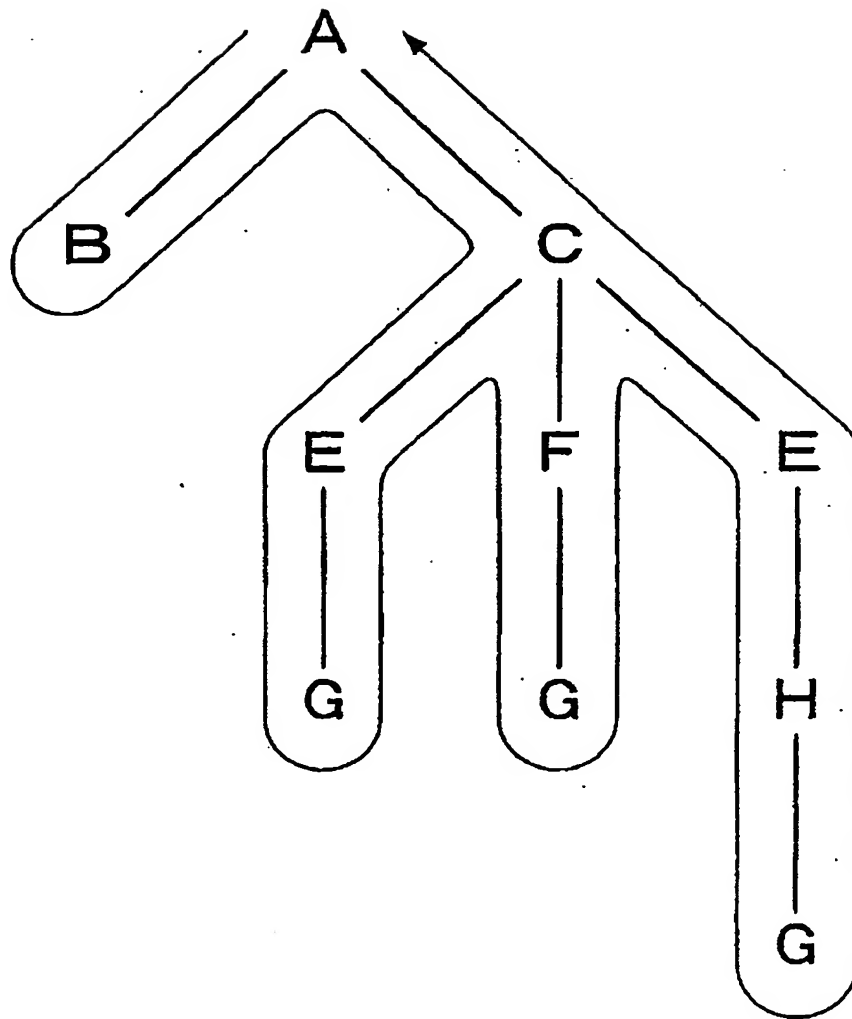


图 8

01:02:08

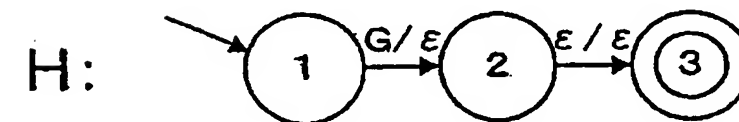
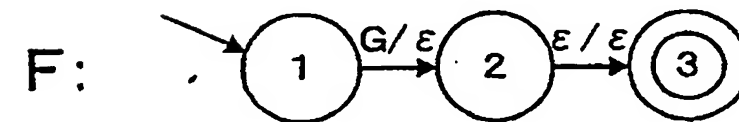
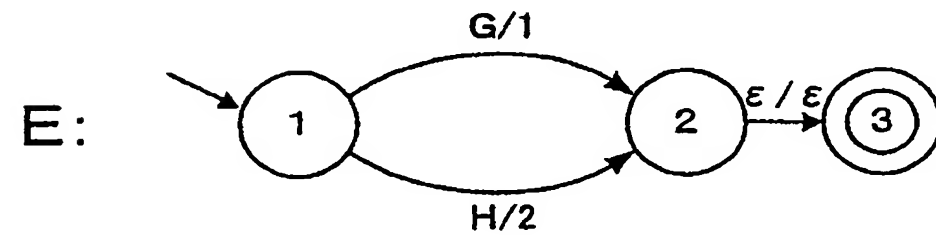
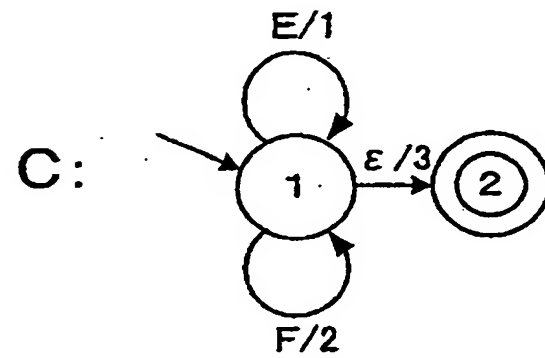
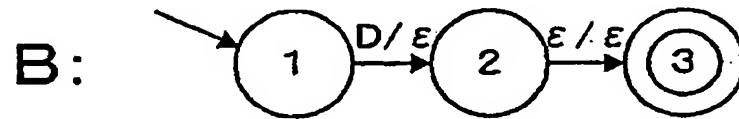
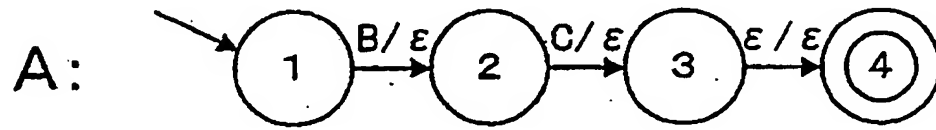


图 9

01.02.08

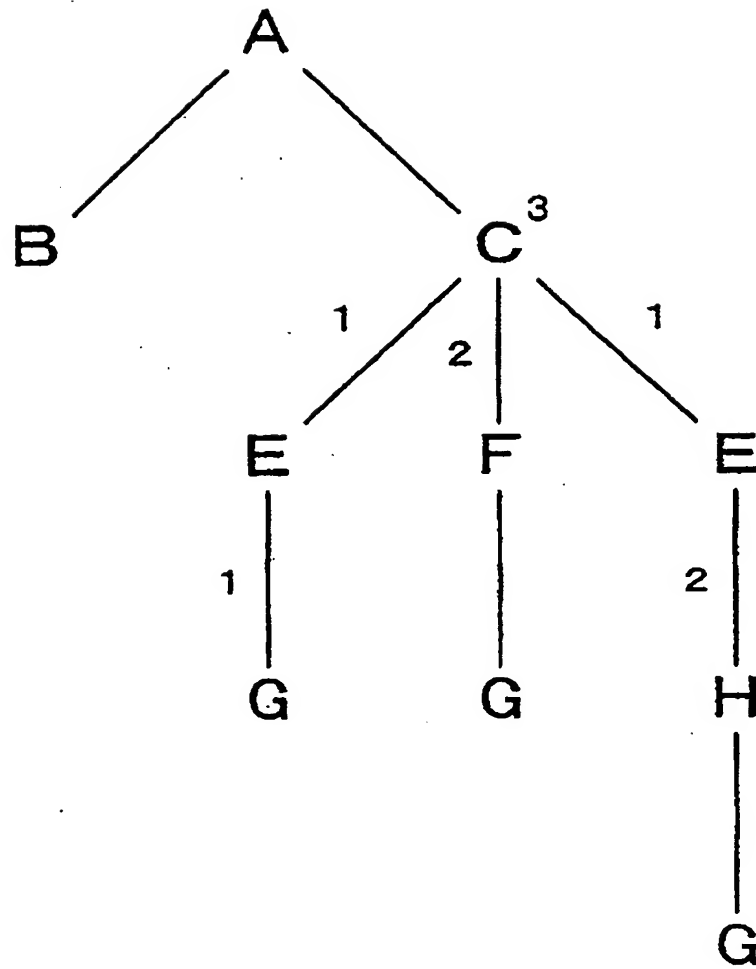


图 10

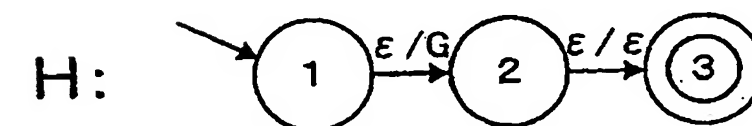
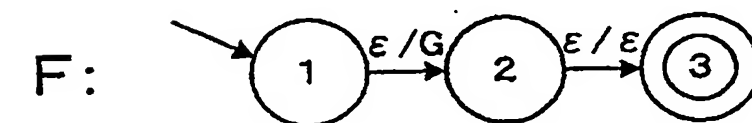
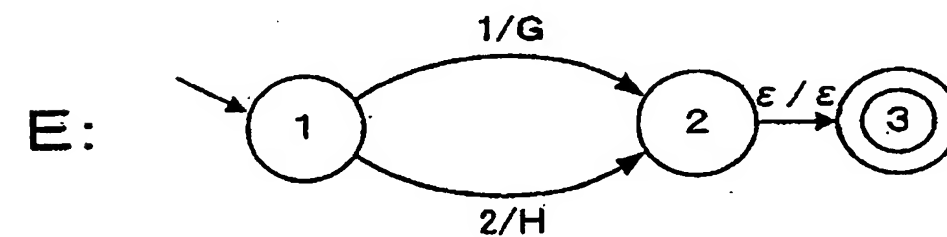
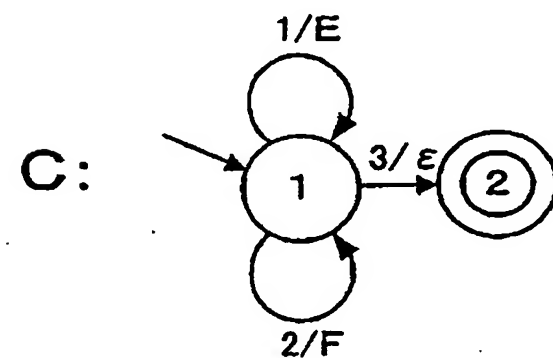
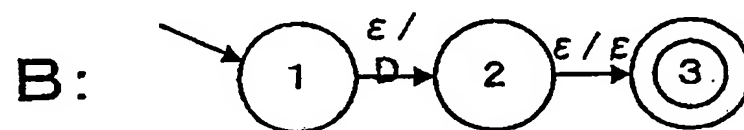
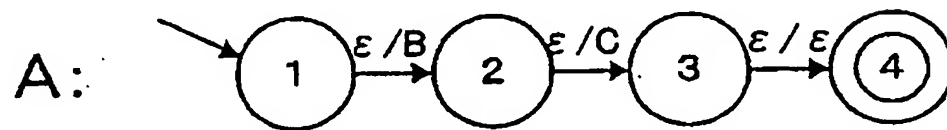


图 11

DTD

转换前	转换后
<pre><!ELEMENT department (employee)> <!ELEMENT employee (name, email)> <!ATTLIST employee serialNo CDATA #REQUIRED> <!ATTLIST employee manager CDATA #IMPLIED> <!ELEMENT name (#PCDATA)> <!ELEMENT email (#PCDATA)></pre>	<pre><!ELEMENT department (employee)> <!ELEMENT employee (serialNo, manager?, name, email)> <!ELEMENT serialNo (#PCDATA)> <!ELEMENT manager (#PCDATA)> <!ELEMENT name (#PCDATA)> <!ELEMENT email (#PCDATA)></pre>

图 12

01.02.08

XML 文件

转换前	转换后
<pre><department serialNo="012345" manager="yes"> <name>aaaaa</name> <email>mail@address.com</email> </department></pre>	<pre><department <serialNo>012345</serialNo> <manager>yes</manager> <name>aaaaa</name> <email>mail@address.com</email> </department></pre>

图 13

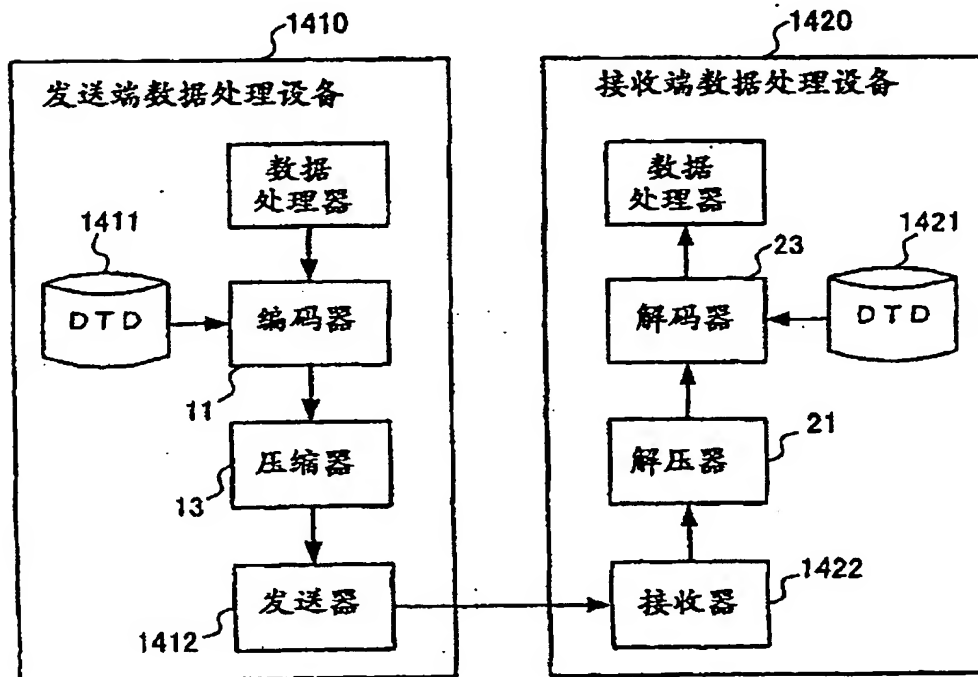


图 14

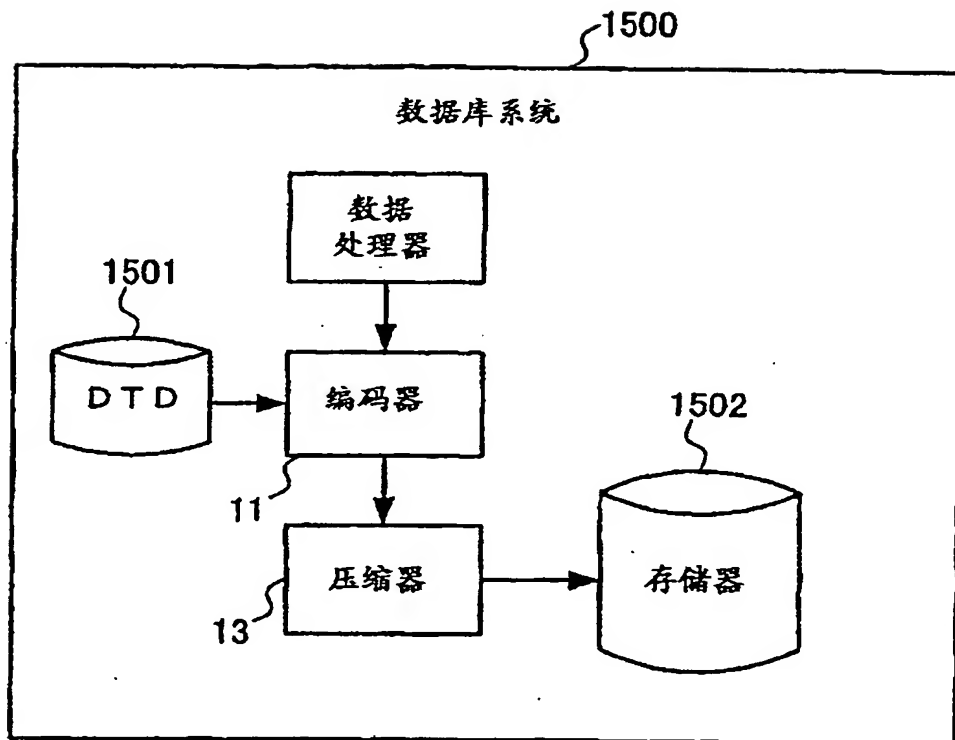


图 15